Day 8: ChIP-seq analysis

Peak calling and scanning for motifs

Rutendo Sigauke
2024

Recap of the videos

1. ChIP-seq introduction
2. Evaluating ChIP-seq data
3. Peak calling with MACS
4. MEME Suite introduction
5. Brief BEDTools introduction
6. ATAC-seq overview (Optional)

Learning Objectives

Downstream analysis of ChIP-seq and ATAC-seq data

- Demonstrate the use of a **peak calling program MACS2** to identify genomic regions with robust signal in each of these data types
  - control/input
  - ENCODE Blacklist
- **Visualize** the raw data and corresponding called peaks
- **Downstream analyses**
  - Motif discovery (MEME)
  - Motif comparison (Tomtom)

Peak calling pipeline

Step 1: Quality control (QC) with fastqc

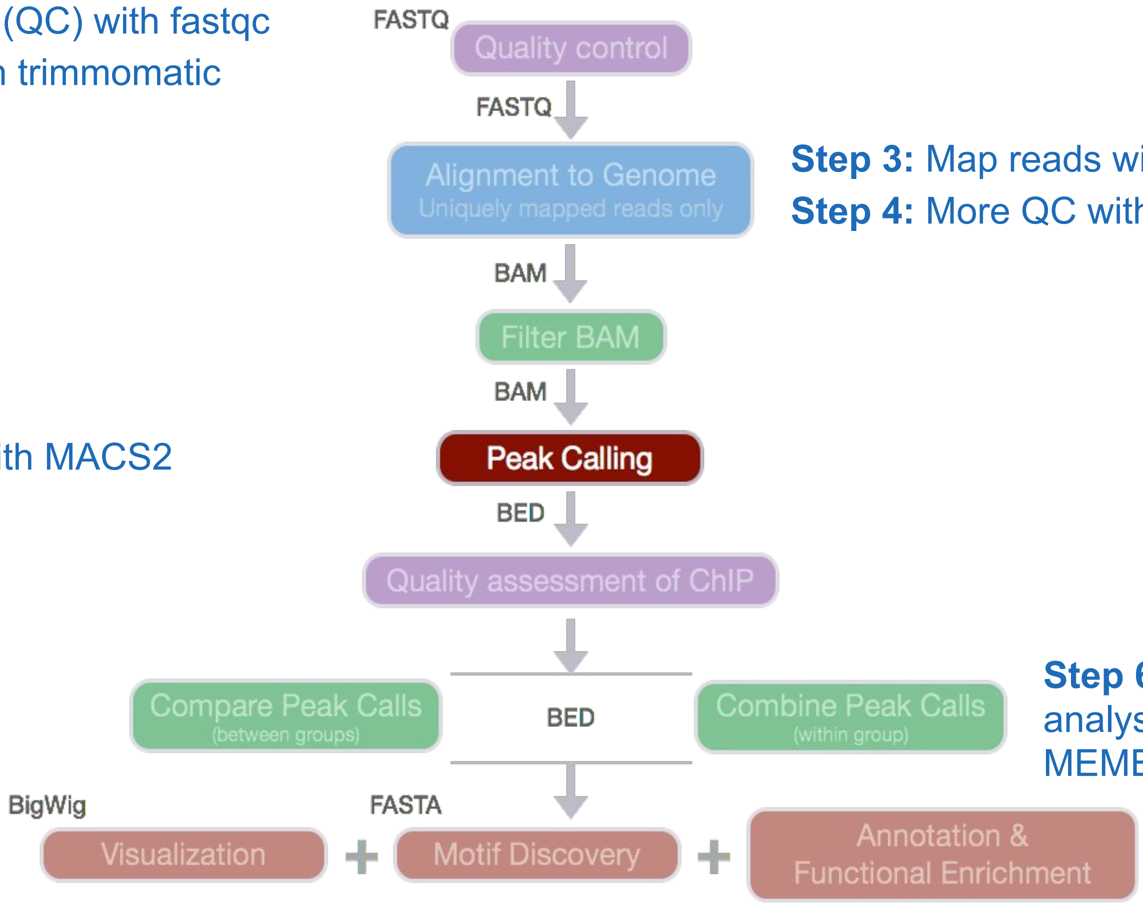
Step 2: Trim reads with trimmomatic

Step 3: Map reads with HISAT2

Step 4: More QC with preseq and MultiQC

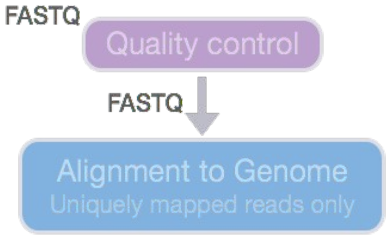
Step 5: Peak calling with MACS2

Step 6: Downstream analyses using bedtools, MEME, TomTom



https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

Steps 1, 2, 3, 4: Quality control and trimming reads



FASTQ

Quality control

FASTQ

Alignment to Genome

Uniquely mapped reads only

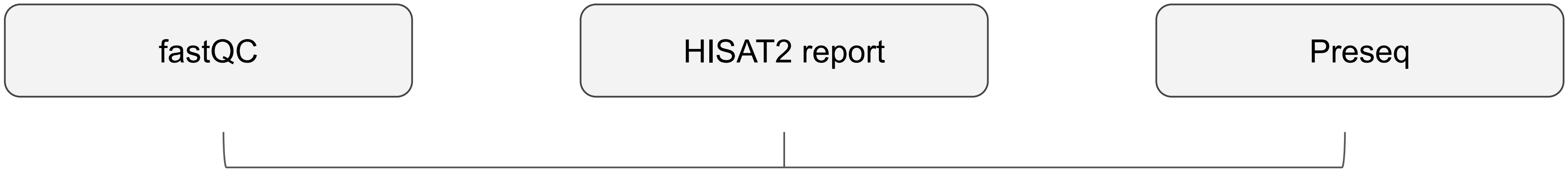
Step 1: Quality control (QC) with fastqc

Step 2: Trim reads with trimmomatic

Step 3: Map reads with HISAT2

Step 4: Quality control via preseq and MultiQC
MultiQC will look through the `qc folder` and create an HTML summary file across all the QC methods

Recap of quality control



fastQC

HISAT2 report

Preseq

Read Sequence Quality

- Base sequence quality
- GC content
- Sequence length and duplication
- Overrepresented sequences

Mapping Summary

Alignment rate per sample

Library Quality

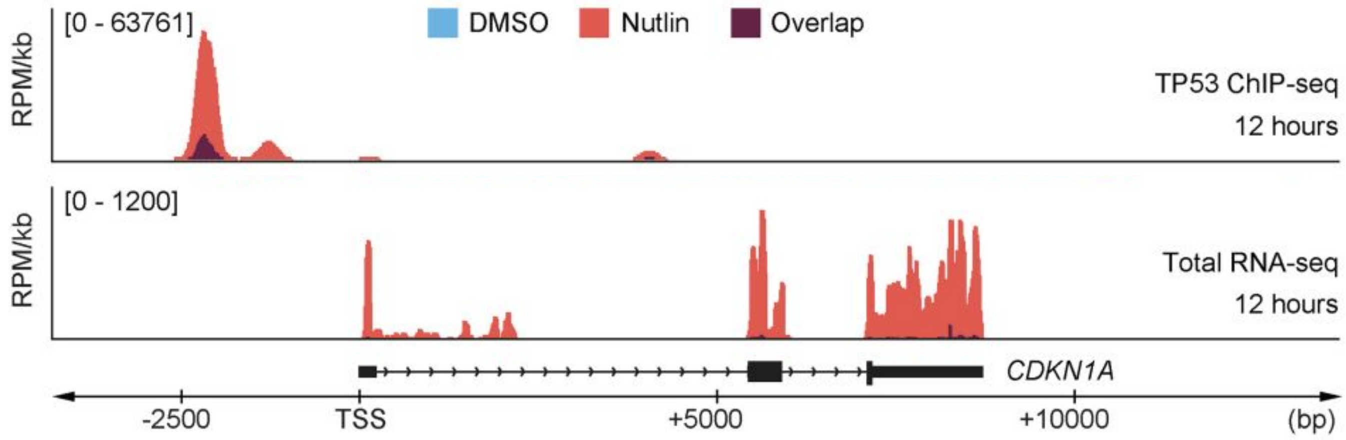
Estimating complexity of library

MultiQC

Map reads to reference genome using HISAT2

ChIP-seq

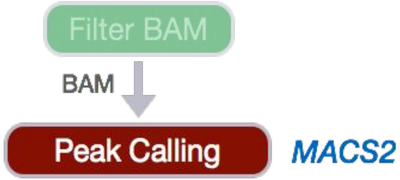
RNA-seq

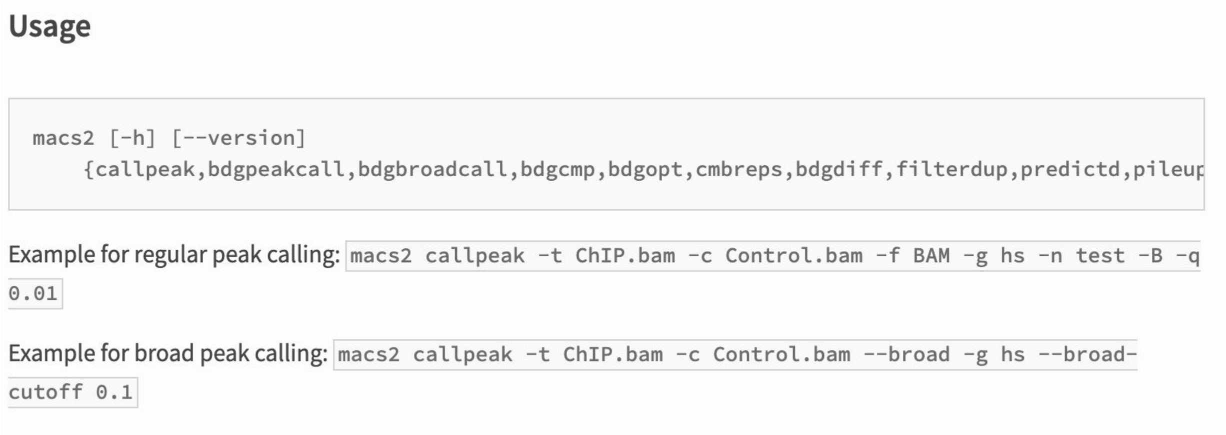


```
hisat2 -p 4 \
        --very-sensitive \
        --no-spliced-alignment \
        -x ${genome} \
        -U ${sample}.trimmed.fastq.gz \
        --new-summary  > ${sam}/${sample}.sam \
        2> ${sample}.hisat2_mapstats.txt
```

Unlike RNA-seq data, there is **no slicing with ChIP-seq data**.

Step 5: Peak calling with MACS2





Usage

```
macs2 [-h] [--version]
    {callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdiff,filterdup,predictd,pileup
```

Example for regular peak calling: `macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01`

Example for broad peak calling: `macs2 callpeak -t ChIP.bam -c Control.bam --broad -g hs --broad-cutoff 0.1`

ChIP-seq peak calling for enrichment

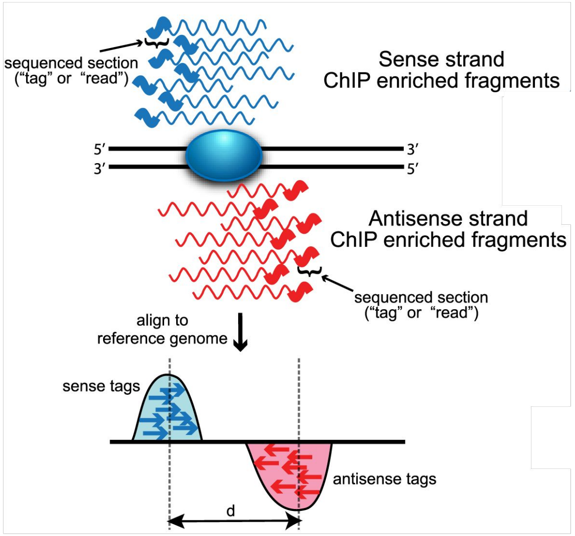


Image source: Wilbanks and Faccioti, PLoS One 2010

ChIP-seq identifies two type of enrichment

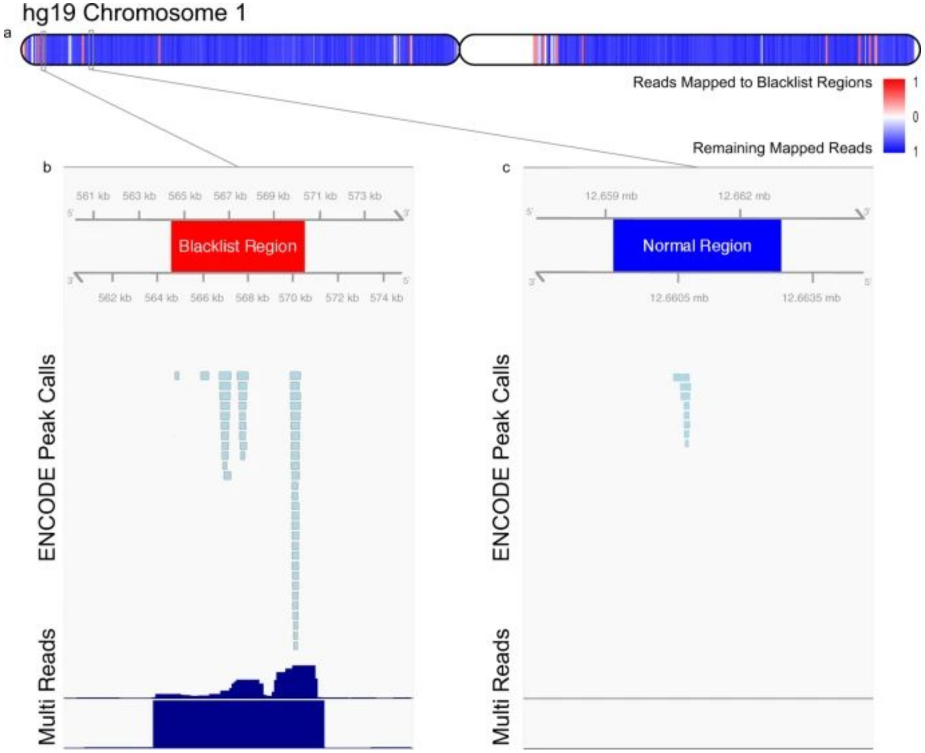
- **Broad peaks:** eg., histone modification. Here we are looking for broad peaks that cover entire gene bodies
- **Narrow peak**: eg., transcription factor binding. Here we are looking for regions of higher amplitude compared to background

MACS genomic input/control

Controls are important!

- ChIP-seq and ATAC-seq are protocols that produce **background noise** as well as **meaningful signal**
  - Therefore, you need controls to not call background noise as peaks
- p/q value cutoffs matter and should vary based on your experiment
- Know your data type: your experiment should inform the parameters of the peak caller
- **Blacklist regions**: some genomic regions almost always show up in these protocols so remove these regions using a Blacklist

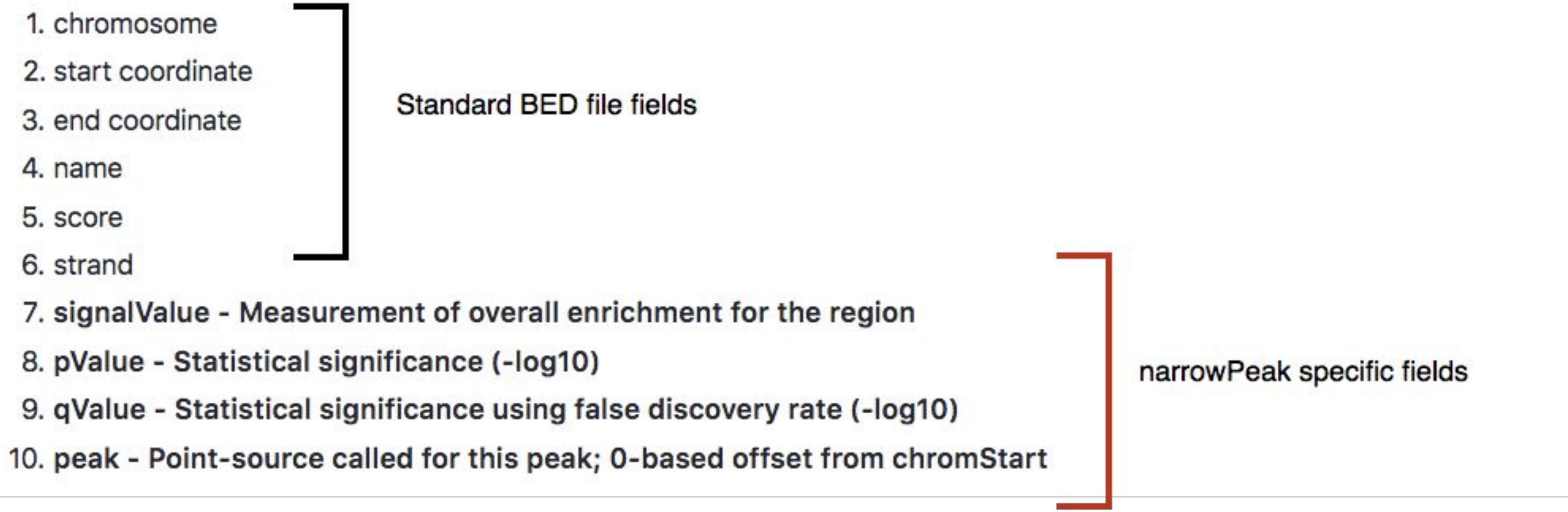
Blacklist regions should be removed



https://www.nature.com/articles/s41598-019-45839-z

These regions contain repetitive regions across the genome and almost always are enriched in ChIP-seq data.

MACS output



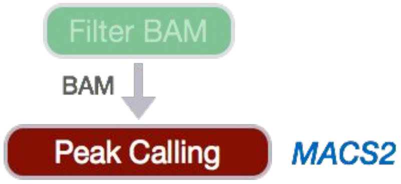
1. chromosome
2. start coordinate
3. end coordinate
4. name
5. score
6. strand
7. **signalValue - Measurement of overall enrichment for the region**
8. **pValue - Statistical significance (-log10)**
9. **qValue - Statistical significance using false discovery rate (-log10)**
10. **peak - Point-source called for this peak; 0-based offset from chromStart**

Standard BED file fields

narrowPeak specific fields

Image: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

Step 5: Peak calling with MACS2

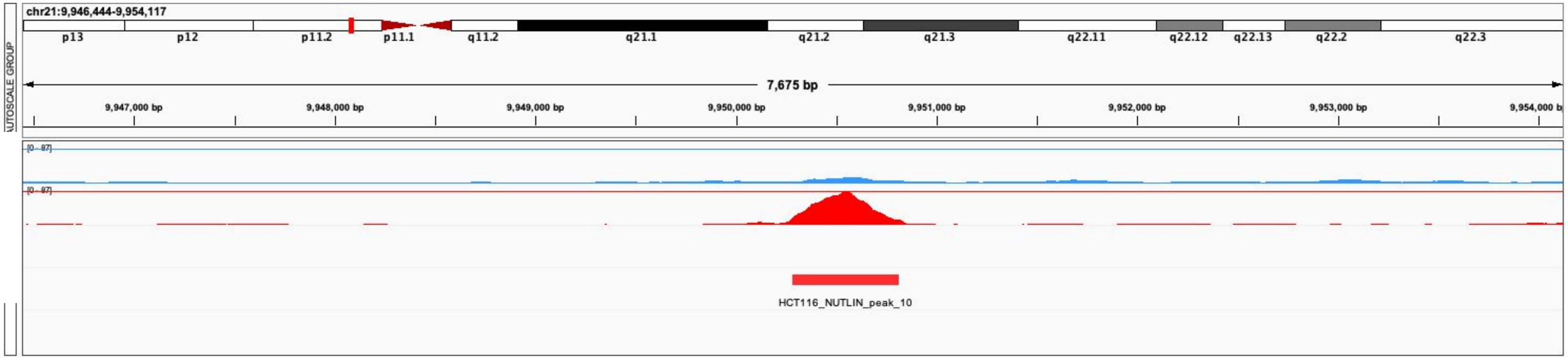


DMSO

NUTLIN

DMSO peak

NUTLIN peak



MACS2 peak calling recommendations

Data type	**`q value`**	**`--broad and --control flags`**	**Reasoning**
ChIP-seq for TF	`<0.01`	`--control <INPUT>`	TF ChIP-seq often has very abrupt, small peaks that are well defined, so narrow peaks is necessary, and a less stringent adjusted p value is likely needed than for other data types

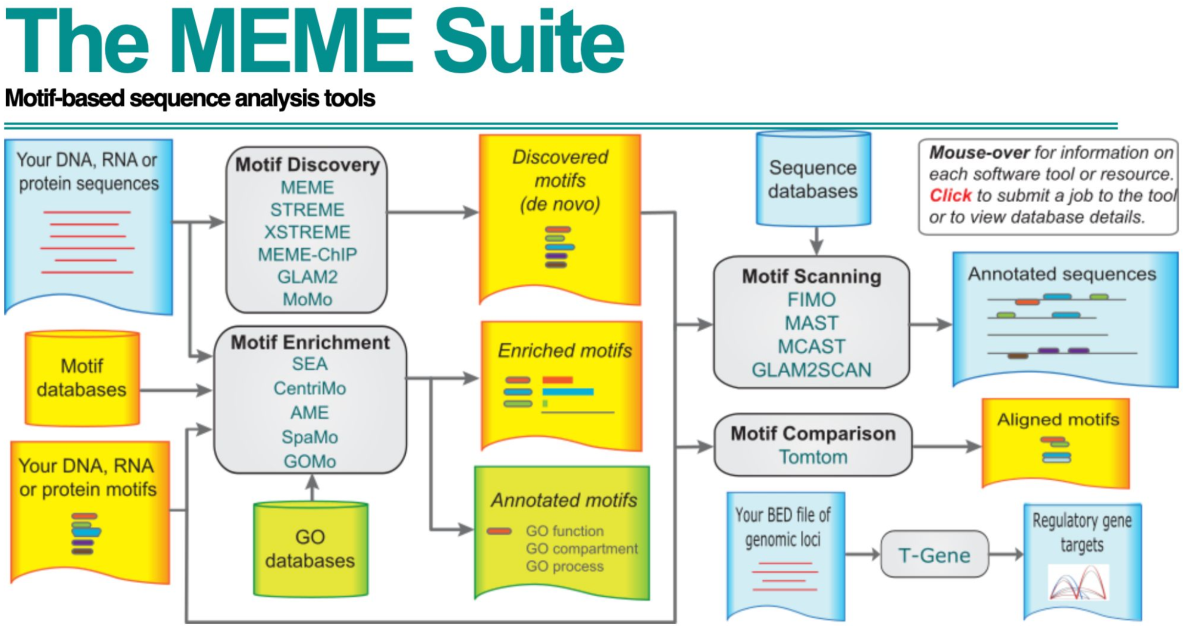
ChIP-seq for histone marks	`<0.0001`	`--broad`
		`--control <INPUT>`
			Histone marks are often broadly dispersed without very well defined edges so a broad peak tag is useful but a very low p value helps differentiate between background and data

ATAC-seq	`<0.0001`	`--control <INPUT>`	ATAC-seq should show peaks at open chromatin across the genome similarly to histone ChIP-seq data, but with more abrupt peaks, so no broad peak tag is needed

Step 6: Downstream analyses with MEME Suite

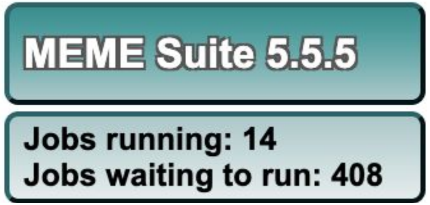
I'M A MAC
AND I'M A PC.
AND SINCE YOU DO EVERYTHING THROUGH A BROWSER NOW, WE'RE PRETTY INDISTINGUISHABLE.

https://xkcd.com/934/



https://meme-suite.org/meme/index.html

We are sharing the MEME Web Server with the world



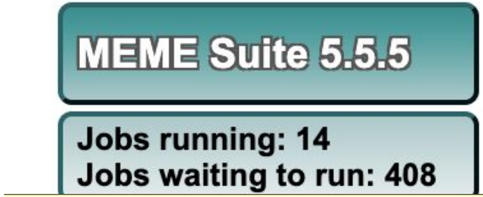
MEME Suite 5.5.5

Jobs running: 14
Jobs waiting to run: 408

1. The web server runs **14 jobs at a time**
2. You can submit up to **4 jobs in an hour**
3. Output from jobs is **retained for 4 days** from the time of submission
4. DO NOT submit any jobs **if over 200 jobs are in the queue**
5. The web server will be **turned off if over 500 jobs** are in the queue

https://meme-suite.org/meme/index.html

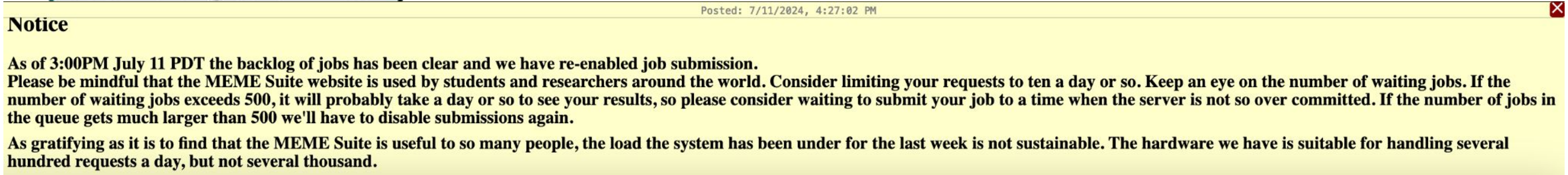
We are sharing the MEME Web Server with the world
Look out for notices



MEME Suite 5.5.5

Jobs running: 14
Jobs waiting to run: 408

1. The web server runs **14 jobs at a time**
2. You can submit up to **4 jobs in an hour**
3. Output from jobs is **retained for 4 days** from the time of submission
4. DO NOT submit any jobs **if over 200 jobs are in the queue**
5. The web server will be **turned off if over 500 jobs** are in the queue



Posted: 7/11/2024, 4:27:02 PM

Notice

As of 3:00PM July 11 PDT the backlog of jobs has been clear and we have re-enabled job submission.
Please be mindful that the MEME Suite website is used by students and researchers around the world. Consider limiting your requests to ten a day or so. Keep an eye on the number of waiting jobs. If the number of waiting jobs exceeds 500, it will probably take a day or so to see your results, so please consider waiting to submit your job to a time when the server is not so over committed. If the number of jobs in the queue gets much larger than 500 we'll have to disable submissions again.

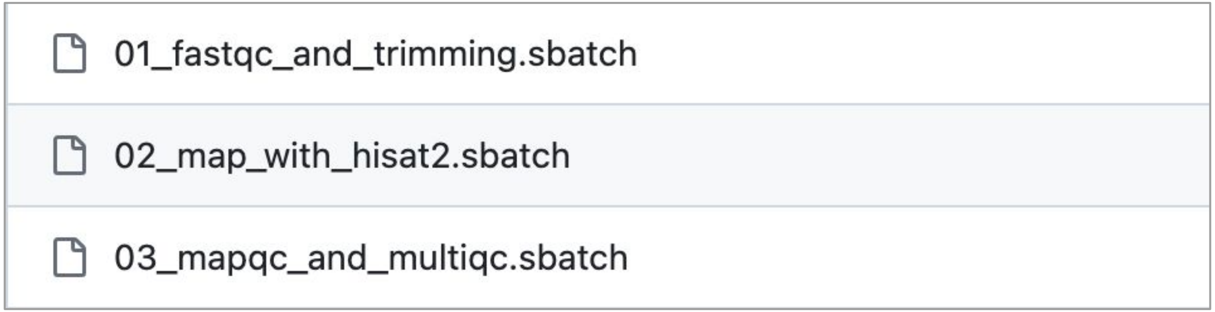
As gratifying as it is to find that the MEME Suite is useful to so many people, the load the system has been under for the last week is not sustainable. The hardware we have is suitable for handling several hundred requests a day, but not several thousand.

https://meme-suite.org/meme/index.html

Processing ChIP-seq Data

The ChIP-seq worksheet (`d8_worksheet1_ChIPseq_analysis.md`) has three steps with 5 scripts.

Section A: Preprocessing of ChIP-seq data

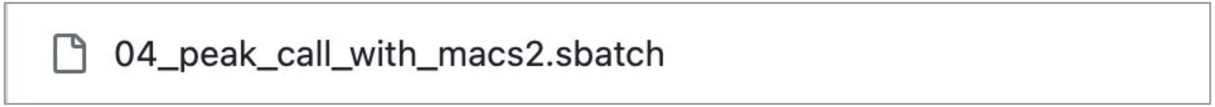


01_fastqc_and_trimming.sbatch

02_map_with_hisat2.sbatch

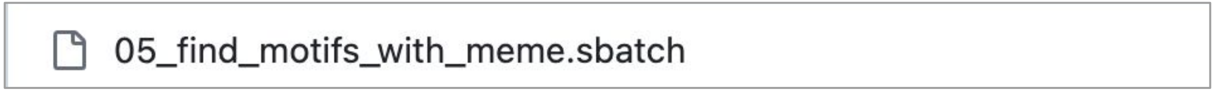
03_mapqc_and_multiqc.sbatch

Section B: Peak calling



04_peak_call_with_macs2.sbatch

Section C: Motif discovery and comparing motifs to database of TF motifs



05_find_motifs_with_meme.sbatch

Additional Resources

Other peak callers:

- Fstitch: https://github.com/Dowell-Lab/FStitch
- SICER: https://zanglab.github.io/SICER2/
- PeakSeq: https://www.nature.com/articles/nbt.1518
- Hpeak: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-369
- PeakRanger: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-139

Other motif discovery tools:

HOMER: http://homer.ucsd.edu/homer/introduction/programs.html